*Volume 18(2), May 2018 http://doi.org/10.17576/gema-2018-1802-13*

# Pembangunan Taksonomi dari Teks Melayu Menggunakan Algoritma Kunang-Kunang Pembahagi Dua Sama

*Mohd Zakree Ahmad Nazri*
*zakree@ukm.edu.my*
*Pusat Teknologi Kecerdasan Buatan, Fakulti Teknologi & Sains Maklumat,*
*Universiti Kebangsaan Malaysia*

*Tri Basuki Kurniawan*
*tribasukikurniawan@gmail.com*
*Pusat Teknologi Kecerdasan Buatan, Fakulti Teknologi & Sains Maklumat,*
*Universiti Kebangsaan Malaysia*

*Abdul Razak Hamdan*
*arh@ukm.edu.my*
*Pusat Teknologi Kecerdasan Buatan, Fakulti Teknologi & Sains Maklumat,*
*Universiti Kebangsaan Malaysia*

*Salwani Abdullah*
*salwani@ukm.edu.my*
*Pusat Teknologi Kecerdasan Buatan, Fakulti Teknologi & Sains Maklumat,*
*Universiti Kebangsaan Malaysia*

*Mohammed Azlan Mis*
*whg5026@ukm.edu.my*
*Fakulti Sains Sosial dan Kemanusiaan,*
*Universiti Kebangsaan Malaysia*

## ABSTRAK

Taksonomi digunakan untuk menerangkan bahawa haiwan boleh dikelaskan kepada beberapa kategori seperti mamalia, reptilia dan buaya. Taksonomi biologi ini membolehkan persamaan, perbezaan malah hubungan antara haiwan ditakrifkan. Konsep dan fungsi taksonomi biologi ini 'dipinjam' oleh saintis dan jurutera Internet dalam membangunkan taksonomi untuk Internet. Seperti taksonomi biologi, membangunkan taksonomi untuk Internet secara manual bukanlah suatu yang mudah dan murah. Tugas ini mengambil masa dan memerlukan kepintaran dalam bidang. Justeru saintis komputer telah menggunakan pendekatan kecerdasan buatan untuk membangunkan taksonomi secara automatik dari teks. Algoritma pembelajaran mesin dicipta untuk membolehkan mesin 'membaca' teks dan kemudiannya 'belajar' untuk membina taksonomi dari konteks yang diperolehi dari teks. Objektif utama kajian ini adalah untuk membangunkan algoritma pembelajaran taksonomi dari Bahasa Melayu yang lebih berkesan dari algoritma sedia ada menggunakan kaedah penghibridan. Makalah ini menyiasat keberkesanan algoritma hibrid antara Algoritma Kunang-Kunang (AKK) dengan Algoritma K-Min Pembahagi Dua Sama (PDS) yang dipanggil Algoritma Kunang-Kunang Pembahagi Dua Sama (AKK-PD). Kajian empirikal ini mengumpul data dari eksperimen yang dijalankan ke atas tiga teks Bahasa Melayu dari bidang Fekah, Biokimia dan Teknologi Maklumat. Perbandingan data ketepatan berasaskan ukuran-F menunjukkan algoritma hybrid AKK-PD membina taksonomi yang lebih tepat berbanding menggunakan algoritma sedia ada. AKK-PD didapati lebih berkesan dan mantap berbanding algoritma bandingan apabila mengendalikan masalah kejarangan data . Walau bagaimanapun, kajian penerokaan ini perlu diteruskan kepada korpus Bahasa Melayu yang lebih besar untuk menguji ketahanan algoritma ini apabila berhadapan dengan korpus yang

eISSN: 2550-2131
ISSN: 1675-8021

lebih umum sifatnya berbanding korpus teks yang teknikal dan menjurus kepada suatu bidang sahaja. Teknik pengekstrakan ciri berasakan kebergantungan sintaksis juga perlu dipertingkatkan kerana jelas teknik telah menghasilkan konteks yang mengalami masalah kejarangan data yang serius. Justeru memberi cabaran baharu untuk penyelidikan pembelajaran taksonomi dari teks Melayu.

**Kata Kunci:** Pembelajaran Mesin; Pembelajaran Taksonomi; Algoritma Kunang-Kunang; Ciri; Teks Bahasa Melayu

# Taxonomy Development from Malay Text Using Firefly Bisection Algorithm

## ABSTRACT

Taxonomy is used to explain that animals can be classified into categories such as mammals, reptiles and crocodiles. This biological taxonomy allows similiarities, differences and relationship between animals to be defined. The concept and function of biological taxonomy is 'borrowed' by Internet scientists and engineers in developing taxonomies for the Internet. Like biological taxonomy, developing taxonomies for the Internet manually is not easy and expensive because the task takes time and requires ingenuity in the field. Thus, computer scientists have used artificial intelligence approaches to develop taxonomies automatically from text. Machine learning algorithms are created to allow the machine to 'read' the text and then 'learn' to construct taxonomy from the context derived from the text. The main objective of this study is to develop an effective taxonomic learning algorithm from Malay text than the existing algorithms using hybridization methods. This study investigates the effectiveness of hybrid algorithms between the Firefly Algorithm (AKK) and the K-Means Bisecting Algorithm (PDS) and thi hybrid algorithm is called the Firefly-Bisecting Algorithm (AKK-PD). This empirical study collects data from experiments carried out on three Malay texts from the Islamic Jurisdiction, Biochemistry and Information Technology. Comparison of accuracy using F-measure shows that the AKK-PD build more accurate taxonomies than using existing algorithms when dealing with data sparseness problem. The AKK-PD is revealed to be more effective and robust compared to the seven existing algorithms. However, this exploratory study needs to be continued with a larger Malay corpus to test the robustness and resilience of this algorithm when dealing with a more general corpus than its technical and specific corpus of texts. The syntactic dependency-based extraction technique needs to be enhanced as it is obvious that this technique have resulted in the context of having serious data sparseness problems. Thus, it opens up new challenge for research about taxonomic learning from Malay texts.

**Keywords:** Machine Learning; Taxonomy Learning; Firefly algorithm; Features; Malay text

## PENGENALAN

Taksonomi adalah satu cadang dari bidang sains untuk pengelasan seperti pengelasan organisme. Guru sains menerangkan hubungan taksonomi antara suatu organisme dengan organisme yang lain. Misalnya, haiwan boleh dikelaskan kepada beberapa kategori seperti mamalia, reptilia dan buaya. Taksonomi biologi ini membentuk suatu skima pengelasan yang membolehkan saintis dan guru mentakrifkan persamaan, perbezaan dan hubungan antara haiwan. Contoh hubungan taksonomi seperti kucing dan gajah adalah mamalia. Konsep dan fungsi taksonomi biologi ini ‘dipinjam’ oleh saintis dan jurutera Internet dalam

membangunkan taksonomi untuk Internet. Taksonomi untuk internet adalah set perkataan dan spesifikasi perkataan seperti jenis hubungan perkataan dengan perkataan lain yang disusun dalam bentuk hierarki yang boleh digunakan untuk memetakan kata kunci dengan sumber yang relevan berasaskan makna atau semantic (Ismail et al., 2009). Seperti taksonomi biologi, membangunkan taksonomi secara manual bukanlah suatu yang mudah dan murah. Tugas ini mengambil masa dan memerlukan pakar bidang untuk membangunkan taksonomi yang berkesan.

Justeru saintis komputer telah menggunakan pendekatan kecerdasan buatan untuk membangunkan taksonomi secara automatik dari teks. Penggunaan pendekatan pembelajaran mesin dalam membangunkan taksonomi secara automatik dipanggil pembelajaran taksonomi. Algoritma pembelajaran taksonomi direka untuk membolehkan mesin 'membaca' teks dan kemudiannya 'belajar' untuk membina taksonomi dari konteks yang diperolehi dari teks. Konteks dalam kajian ini bermaksud memeroleh ciri-ciri iaitu 'kata kerja' yang sering digunakan bersama dengan 'kata nama'. Untuk memeroleh ciri ini, kaedah pemerolehan ciri berasaskan kebergantungan sintaksis digunakan. Tetapi menggunakan kaedah ini boleh menyebabkan masalah kejarangan data. Masalah kejarangan data bermaksud, ciri (kata kerja) untuk suatu kata nama yang menjadi asas konteks agar kata nama tersebut dapat ditentukan kedudukannya dalam suatu 'hierarki konsep' adalah tidak mencukupi.

Oleh yang demikian, wujud keperluan membangunkan algoritma pembelajaran taksonomi yang mantap iaitu berupaya menjana taksonomi yang berkualiti walau pun data yang diperolehi mengalami masalah kejarangan yang serius. Makalah ini membincangkan dengan terperinci bagaimana AKK dapat membentuk taksonomi dari teks Melayu dan melaporkan keberkesanan algoritma hibrid antara Algoritma Kunang-Kunang (AKK) dengan Algoritma K-Min Pembahagi Dua Sama (PDS) yang dipanggil Algoritma Kunang-Kunang Pembahagi Dua Sama (AKK-PD).

Kertas ini dimulakan dengan pengenalan dan kajian lepas. Bahagian seterusnya menjelaskan metodologi kajian yang digunakan dalam kajian ini. Bahagian ini turut membincangkan tetapan eksperimen. Teknik yang dibangunkan dijelaskan secara terperinci dalam bahagian ketiga. Selepas algoritma cadangan dibincangkan, keputusan eksperimen akan dibentangkan. Bahagian keempat melaporkan hasil analisis keputusan dan perbandingan keputusan dengan algoritma lain. Penilaian prestasi teknik tersebut dilaporkan dalam bahagian ini. Bahagian keenam pula membincangkan hasil kajian ini dan makalah ini diakhiri dengan kesimpulan keseluruhan kajian ini.

## SOROTAN KAJIAN PEMBELAJARAN TAKSONOMI DARI TEKS

Penyelidikan terkini seperti Ristoski et al. (2017), Cimiano (2016) dan Anh et al. (2017) mencadangkan teknik baharu dalam membangunkan taksonomi secara automatik dari teks Bahasa Inggeris. Namun kaedah dan teknik yang dibangunkan hanya tertumpu kepada teks berbahasa Inggeris. Wang et al. (2017) menyatakan bahawa penyelidikan pembelajaran taksonomi dari teks selain dari Bahasa Inggeris adalah suatu yang bernilai untuk diterokai kerana struktur dan gaya bahasa yang berbeza. Persoalan kajian selalunya berkisar kepada pengujian teknik pembelajaran mesin yang sebelum ini terbukti berkesan pada teks Inggeris. Namun adakah teknik tersebut turut berkesan kepada teks bahasa lain seperti Bahasa Melayu? Antara isu yang sering menghambat pembelajaran taksonomi adalah isu kejarangan data. Kejarangan data boleh disebabkan oleh banyak faktor, antaranya adalah kaedah pengekstrakan fitur (perkataan) dari teks (korpus).

Cimiano (2006) menyatakan bahawa nilai atribut bagi suatu istilah yang diekstrak kemungkinan boleh jadi salah atau mengalami masalah kejarangan kerana perkara berikut: (i) alat pemproses bahasa tabie gagal melabel jenis kata dengan tepat. Oleh itu, tidak semua kata

kerja yang diektrak berasaskan kebergantungan sintaksis adalah betul; (ii) tidak semua kebergantungan sintaksis yang diperolehi (walau pun betul) akan membantu untuk membezakan antara suatu objek dengan objek yang berlainan; dan (iii) andaian kesempurnaan maklumat tidak akan pernah dapat dipenuhi (Zipf, 1935).

Kajian dalam bidang ini sering tertumpu kepada pembangunan algoritma pembelajaran taksonomi dari teks yang lebih berkesan dari algoritma sedia ada. Beberapa kajian seperti Cimiano (2005, 2006) dan de Mantaras dan Saitia (2004) telah menggunakan algoritma seperti Analisis Konsep Formal, GAHC, pengelompokan agglomerat dan K-Min Pembahagi Dua Sama untuk menyelesaikan isu hingar dan kejarangan data. De Castro dan Timmis (2002b) menggambarkan beberapa ciri-ciri yang wajar dimiliki oleh suatu algoritma dan satu daripadanya ialah keteguhan atau ketahanan algoritma untuk tetap menghasilkan taksonomi yang berkualiti apabila 'dihidangkan' dengan data yang bermasalah. Menurut Wang et al. (2017), kaedah pembelajaran taksonomi dari teks boleh dibahagikan kepada tiga kategori iaitu:

1) pengekstrakan hubungan taksonomi berasaskan pola Bahasa (Nazri et al., 2008)
2) kaedah berasaskan pengagihan (ciri)
3) pembangunan taksonomi dari ayat yang mengandungi kata kunci taksonomi iaitu '*adalah*'. Sebagai contoh, '*kucing adalah mamalia'*.

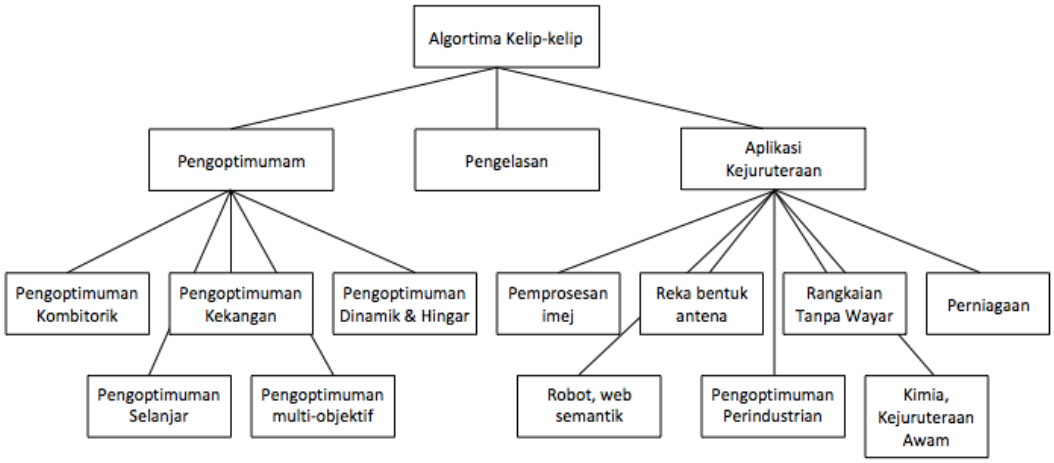
Kajian ini tertumpu kepada kategori ke dua iaitu pembelajaran taksonomi berasaskan pengagihan. Kaedah ini tergolong dalam kaedah pembelajaran mesin tanpa selia. Kaedah ini adalah berdasarkan kepada hipotesis pengagihan Harris (1954) untuk mereka bentuk teknik untuk mengenalpasti sinonim, konsep dan hubungan taksonomi. Harris (1954) menyatakan bahawa 'perkataan adalah sama sekiranya ia berkongsi konteks yang serupa'. Firth (1957) memperkenalkan sifat kebergantungan-konteks dengan idea 'situasi konteks'. Kenyataan Firth (1957) bahawa *'anda akan mengetahui suatu perkataan melalui perkataan lain yang bersamanya'* telah menjadi satu faktor penting dalam penyelidikan perlombongan teks, capaian maklumat dan pembelajaran ontologi. Oleh yang demikian, teori yang mendasari kajian ini adalah seperti berikut:

1) Kontekstual (pengagihan) hipotesis makna (Harris 1954) dan (Firth, 1957) iaitu makna perkataan bergantung kepada penggunaannya dalam teks.
2) Hipotesis konteks persamaan semantik (Miller dan Charles 1991), perkataan yang mempunyai persamaan konteks menggambarkan ia turut mempunyai persamaan semantik.

Data yang dikumpul oleh Charles (2000) mengesahkan dakwaan bahawa manusia mengikhtisar perwakilan konteks dari pengalaman beberapa konteks linguistik suatu perkataan. Penemuan ini menyokong hipotesis konteks makna. Terdapat beberapa penyelidikan telah dijalankan untuk memeriksa kesahihan hipotesis seperti Jiang dan Conrath (1997) dan Charles (2000). Siasatan empirikal mereka telah mengesahkan kesahihan hipotesis di atas.

Kebanyakan kajian yang menggunakan kaedah ini menggunakan ukuran persamaan taburan seperti kosin, Jaccard, Jensen-Shannon divergience atau ukuran LIN untuk mengukur hubungan antara suatu perkataan ketika membentuk hierarki konsep. Bafna dan Wiens (2015) menggunakan ukuran persamaan bukan Euclidean manakala Lefever (2016) menggunakan rangkaian neural buatan berasaskan hipotesis pengagihan dan Kang (2016) pula menggunakan kaedah berasaskan graf. Kajian kesusasteraan menunjukkan kaedah ini masih terus digunakan seperti Wang et al. (2018) yang mereka kaedah pembelajaran taksonomi untuk teks Bahasa Cina. Beberapa kajian terdahulu yang dijalankan seperti Cimiano (2005, 2006) dan de Mantaras dan Saitia (2004) tertumpu kepada penyelesaian isu hingar dan

kejarangan data. De Castro dan Timmis (2002) menggambarkan beberapa ciri-ciri yang wajar dimiliki oleh suatu algoritma dan satu daripadanya ialah keteguhan atau ketahanan algoritma untuk tetap menghasilkan taksonomi yang berkualiti apabila 'dihidangkan' dengan data yang bermasalah.




RAJAH 1. Aplikasi Algoritma Kunang-Kunang

Beberapa variasi algoritma kunang-kunang wujud dalam kesusasteraan. Fister et al. (2012) telah mencadangkan skim klasifikasi untuk mengklasifikasikan AKK kepada beberapa kategori dengan berasaskan kepada tetapan parameter mereka. Pengaturan tetapan parameter AKK ini adalah penting untuk mencapai prestasi yang baik yang perlu dipilih dengan berhati-hati. Secara umum, terdapat dua cara untuk menetapkan parameter algoritma dengan betul. Pertama adalah melalui penalaan nilai parameter sebelum algoritma dilarikan dan kaedah kedua adalah penalaan parameter dijalankan setelah AKK dilarikan. Penalaan parameter dijalankan setelah lengkap suatu lelaran. Selain daripada berdasarkan kaedah penetapan parameter, pengelasan yang digunakan oleh Fister et al. (2012) juga mengambil kira komponen dan ciri apakah yang terdapat algoritma AKK berkenaan.

Tidak dinafikan bahawa AKK telah direkayasa untuk menjadi sebuah algoritma pengelompokan. Senthilnath et al. (2011) telah membandingkan prestasi algoritma pengelompokan yang diinspirasi bio, termasuk AKK, Koloni Lebah Buatan (ABC) dan PSO (Pengoptimuman Partikel Kerumunan) dan menyimpulkan bahawa AKK lebih cekap untuk pengelompokan data. Sarma dan Gopi (2014) telah menggunakan AKK untuk proses pengelompokan dalam penyelidikan berkaitan rangkaian penderia tanpa wayar (WSN). Wong et al. (2014) telah menggunakan AKK untuk membolehkan sistem storan tenaga menyimpan tenaga elektrik secara optimal. Mohammed et al. (2015) dalam penelitiannya mengelompokkan secara berhierarki dokumen menggunakan keadah AKK, dimana setiap kunang-kunang mewakili satu dokumen, dengan urutan proses pra-pemrosesan dokumen, pengelompokan dokumen dan pengelompokan kembali (*re-locating*) dokumen, membentuk susunan dokumen mengikut relevansi kepada isi dari dokumen tersebut. Apa yang dapat disimpulkan dari Rajah 1 yang terhasil di atas adalah pembelajaran ontologi dan pembelajaran taksonomi sama ada dari teks Inggeris atau pun Melayu masih belum diuji dengan AKK atau pun yang diinspirasikan oleh AKK.

Algoritma Kunang-Kunang (AKK) yang dibangunkan oleh Yang (2008) menunjukkan ketahanan dan kecekapan dalam menyelesaikan masalah rekabentuk tidak linear yang merupakan suatu masalah polynomial tak berkententuan. Menurut Fister et al. (2013), untuk menggunakan AKK bagi menyelesaikan pelbagai masalah, AKK perlu diubahsuai atau dihibrid dengan algoritma lain. Saranan Fister et al. (2013) terbukti berkesan

apabila beberapa penyelidik lain data seperti Jain et al. (2017) telah menggunakan AKK dan Set Kasar untuk mengelompok imej. Hassanzadeh dan Meybodi (2012) telah menghibrid AKK dan K-means manakala Lei et al. (2016) menggunakan AKK untuk mengenalpasti protein. Nayak et al. (2014) telah menghibrid AKK dengan C-Min Kabur untuk mengelompok data dan Mohammed et al. (2014) menggunakan AKK untuk mengelompok dokumen setelah AKK diubahsuai. Rajah 1 menunjukkan taksonomi aplikasi AKK dalam beberapa domain. Berdasarkan kajian kesusasteraan yang dijalankan, maka bolehlah disimpulkan bahawa AKK belum digunakan untuk tugas pengelompokan berhirarki untuk pembelajaran taksonomi kerana AKK sahaja tidak dapat membentuk taksonomi dan AKK perlu dihybrid untuk mencapai objektif pembelajaran taksonomi. Oleh yang demikian makalah ini mencadangkan agar AKK dihibrid dengan algoritma yang boleh membentuk hierarki (taksonomi) iaitu Algoritma K-Min Pembahagi Dua Sama (PDS).

## ALGORITMA KUNANG-KUNANG

Menurut Lewis dan Cratsley (2008), serangga kelip-kelip atau kunang-kunang (Coleoptera: Lampyridae), adalah antara serangga yang paling 'berkarisma' dikalangan serangga, terutamanya cara kunang-kunang memikat yang telah memberi inspirasi kepada saintis. Menurut Hudawiyah et al. (2015), terdapat perbezaan yang ketara pada sistem saraf kunang-kunang jantan dan betina yang menunjukkan bahawa seranga ini mempunyai fungsi kawalan yang berbeza antara kunang-kunang jantan dan betina. Kunang-kunang adalah sejenis serangga dari kumpulan kumbang. Menurut Fister et al. (2013), terdapat lebih 2000 spesis kunang-kunang di dunia dan spesis 'Pteroptyx Tener' adalah spesis yang terdapat di Kuala Selangor dan Sungai Sepetang di Kampung Dew, Perak (Hazmi & Sagaff, 2017; Juliana et al., 2012; Jusoh et al., 2013). Keunikan yang ada pada kunang-kunang ini adalah pada ekornya, yang mengeluarkan kelipan cahaya. Serangga kunang-kunang lebih unik kerana cahaya yang dihasilkan oleh kerumunan serangga ini adalah serentak iaitu 3 kelipan sesaat.
Kunang-kunang hanya 6 mm panjang hidup dalam iklim tropika di kawasan berpaya kerana sumber makanan utamanya adalah dari pokok Berembang. Pokok Berembang atau 'Sonneratia Caseolaris', sejenis pokok paya yang tumbuh liar adalah sebahagian dari ekosistem penting bagi kunang-kunang. Selain daripada itu, pokok Berembang juga penting sebagai habitat yang bertindak sebagai penapis untuk kotoran dan racun dan mengeluarkan air bersih untuk organisma dalam sungai. Jangkamasa hayat kunang-kunang ialah selama 2 hingga 3 bulan.

Kelipan yang dihasilkan oleh kunang-kunang adalah hasil proses biokimia iaitu biopendarcahaya *(bioluminescence)*. Organ yang menghasilkan tindak balas biopendarcahaya yang menghasilkan cahaya ialah lantera. Kunang-kunang jantan mengeluarkan cahaya yang lebih terang berbanding yang betina untuk menarik perhatian kunang-kunang betina. Kunang-kunang jantan berupaya mengawal biopendarcahaya untuk menghasilkan cahaya yang kuat dan berlainan (unik). Selain isyarat memikat, kelipan cahaya juga adalah untuk memberi amaran dan menghalau pemangsa. Namun ada juga kunang-kunang yang tidak mampu menghasilkan tindak balas biopendarcahaya. Kunang-kunang ini akan memikat pasangannya dengan menghasilkan feromon seperti semut. Lantera penghasil kelipan cahaya ini dimulakan oleh isyarat yang dikeluarkan dari sistem saraf pusat kunang-kunang.

Kebanyakan kunang-kunang bergantung kepada biopendarcahaya untuk memikat pasangannya. Lazimnya, kunang-kunang jantan akan mengeluarkan isyarat pertama untuk memikat kunang-kunang betina. Kunang-kunang betina akan membalas isyarat kunang-kunang jantan dengan menghasilkan kelipan cahaya yang berterusan. Kedua-dua kunang-kunang jantan dan betina menghasilkan isyarat cahaya yang pola kelipannya berbeza dan pemasaannya tepat untuk mengekod maklumat seperti identiti spesis dan jantina. Kunang-

kunang betina tertarik kepada kunang-kunang jantan yang menghasilkan isyarat mengawan yang menunjukkan perbezaan sifat atau unik. Kebiasaannya, kunang-kunang betina tertarik kepada kelipan cahaya yang lebih terang.

Walau bagaimana pun, aras kecerahan cahaya atau keamatan cahaya berbeza-beza mengikut jarak sumber dan kunang-kunang betina tidak dapat membezakan di antara keamatan cahaya yang disebabkan oleh jarak atau kekuatan cahaya yang dihasilkan oleh lantera kunang-kunang jantan. Pancaran cahaya kunang-kunang amat mudah dilihat dan berkesan sebagai mekanisme pertahanan yang memberi amaran kepada pemangsa dan sekali gus mengelak kunang-kunang dari menjadi mangsa. Terdapat dua ciri kecerdasan kerumunan iaitu i) organisasi swaatur dan ii) pembuatan keputusan tak terpusat. Kehidupan sosial kunang-kunang adalah sama seperti serangga lain yang hidup sekawan atau berkumpulan seperti lebah dan semut. Setiap kunang-kunang, seperti semut adalah individu berautonomi yang hidup dalam satu koloni yang harmoni. Keharmonian dicapai apabila setiap ekor kunang-kunang tidak hidup terpencil tetapi hidup bersosial iaitu saling berinteraksi dan berkomunikasi di antara satu sama lain. Bagi kunang-kunang, kehidupan sosial atau kemasyarakatan ini bukan sahaja dalam aktiviti mencari makanan tetapi juga dalam pembiakan. Pembuatan keputusan secara kolektif berkait rapat dengan sifat dan tingkah laku memancarkan cahaya yang menjadi asas dan inspirasi kepada pembangunan algoritma kunang-kunang oleh saintis komputer bernama (Yang, 2010).

Algoritma 1 menunjukkan bagaimana AKK berfungsi. Yang (2010) mereka bentuk AKK dengan suatu andaian bahawa semua Kunang-Kunang adalah uniseks agar mana-mana Kunang-Kunang akan tertarik untuk mendekati kesemua Kunang-Kunang yang lain. Tarikan adalah berkadar dengan kecerahan mereka. Kunang-Kunang yang kurang terang kelipannya akan tertarik (dan dengan itu bergerak ke arah) Kunang-Kunang yang lebih cerah. Bagaimanapun, keamatan (aras kecerahan cahaya) akan berkurangan dengan peningkatan jarak di antara mereka. Jika tidak ada Kunang-Kunang yang lebih terang dari Kunang-Kunang yang saat ini diamati, Kunang-Kunang berkenaan akan bergerak secara rawak. Kecerahan atau keamatan harus dikaitkan dengan fungsi objektif. Algoritma 1 adalah pseudokod algoritma Kunang-Kunang asal yang dibangunkan oleh Yang (2010).

1: $t = 0; s *= \varnothing;\ \gamma = 1.0;$ *// mengasal: bilLelaran, solusi terbaik, dayaPenarik*
2: $P^{(0)} = janaPopulasiAwal(\ \ )$; //jana populasi awal
3: $Sementara\ (t < MAKSFES)\ lakukan$
4: $a^{(t)} = AlphaBaru()$ // menentukan nilai $\alpha$ yang baharu
5: $NilaiAKK\left(P^t, f(s)\right)$; //nilai *s* berdasarkan kepada $f(s)$
6: $Isi\left(P^t, f(s)\right)$; // isih *s* berdasarkan kepada $f(s)$
7: $S^* = CariKelipKelipTerbaik\left(P^t, f(s)\right)$ // tentukan solusi terbaik
8: $P^{t+1} = GerakAKK(P^{t+1})$ //pelbagaikan daya tarikan sewajarnya
9: t=t+1;
10: Lelaran Tamat

ALGORITMA 1. Pseudokod Algoritma Kunang-Kunang Yang Dibangunkan Oleh Yang (2010)

Algoritma Kunang-Kunang adalah berasaskan kepada rumus keamatan cahaya I yang berkurangan dengan bertambahnya jarak $r^2$. Walau bagaimanapun, sekiranya jarak dari sumber cahaya bertambah, penyerapan cahaya menyebabkan cahaya menjadi semakin lemah. Fenomena ini boleh dikaitkan dengan fungsi objektif yang perlu dioptimumkan. Daya tarikan adalah berkadar kepada keamatan cahaya. Keamatan cahaya mempunyai kesan atau ditentukan oleh landskap fungsi kesesuaian. Populasi Kunang-Kunang diasalkan oleh fungsi '*mengasalAKK*'. Lazimnya proses mengasal ini dijalankan secara rawak. Proses gelintaran

AKK berada di gelung sementara (baris 3 – 10 dalam Algoritma 1 Rajah 4.1). Fungsi *AlphaBaru* adalah untuk mengubahsuai nilai parameter $\alpha$ asalan. Fungsi *NilaiAKK* pula menilai kualiti solusi. Fungsi kesesuaian *f(s)* dilaksanakan di dalam fungsi ini. Fungsi *Isih* mengisih populasi Kunang-Kunang berdasarkan kepada nilai fungsi kesesuaian. Fungsi *CariKelipKelipTerbaik* memilih individu Kunang-Kunang terbaik daru populasi. Fungsi *GerakAKK* pula menentukan kedudukan Kunang-Kunang dalam ruang gelintaran iaitu Kunang-Kunang digerakkan kepada individu yang menarik. Gelintaran Kunang-Kunang dikawal oleh jumlah maksimum fungsi kesesuaian (MAKSFES).

## ALGORITMA KUNANG-KUNANG PEMBAHAGI DUA SAMA (AKK-PD)

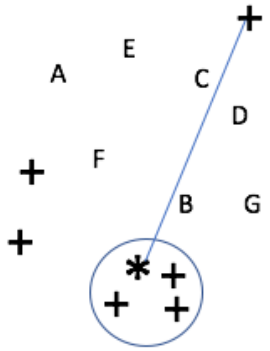
Bahagian ini membincangkan rekabentuk algoritma hibrid AKK dan K-Min Pembahagi Dua Sama yang dinamakan sebagai Algoritma Kunang-Kunang Pembahagi Dua Sama atau AKK-PD. Perbezaan utama antara AKK-PD dan AKK asal adalah pembentukan hierarki kelompok AKK-PD dilaksanakan oleh algoritm K-Min Pembahagi Dua Sama kerana AKK asal tidak direka untuk membentuk hierarki konsep atau taksonomi. Algoritma 2 berikut memaparkan pseudokod AKK-PD.

1: Input: data (kata nama dan ciri) yang diekstrak dari teks.
2: *Mengasal: bilLelaran, solusi terbaik, dayaPenarik, nilai ambang pengelompokan* ($t = 100; s * = \varnothing;\ \gamma = 1.0; p = 0.05; bilPopulasi = 50$)
3: $P^{(0)} = populasiAwalKelipKelip(\ \ );$
Jana satu populasi Kunang-Kunang sebanyak 50 ekor. Seekor Kunang-Kunang mempunyai sebanyak Kmax sentroid. Kmax asal = jumlah kata nama yang akan dikelompokkan. Setiap k-ekor Kunang-Kunang adalah satu solusi.
4: While (lelaran <=t)
5: Untuk Setiap Ekor Kunang-Kunang
Lokasi setiap sentroid asal adalah sama dengan lokasi data
6: Untuk Setiap Sentroid dalam Kunang-Kunang
7: Ukur jarak dengan sentroid lain dalam Kunang-Kunang yang sama
8: Sentroid yang mempunyai jarak yang kurang dari (p * jarak_terjauh) akan digabungkan dengan memilih sentroid baharu sebagai kedudukan baharu.
9: Untuk semua data
10: Ukur jarak data dengan semua sentroid
11: Kelompokkan data kepada sentroid yang terdekat (menggunakan persamaan kesamaan ciri)
12: Kira kualiti setiap setiap kelompok
13: Gerakkan Kunang-Kunang kepada Kunang-Kunang yang mempunyai nilai *i* yang paling tinggi. Pergerakan Kunang-Kunang menuju ke arah Kunang-Kunang yang memiliki i yang tertinggi ditentukan dengan menggunakan rumus X.
14: Lelaran++
15: Ulang langkah di baris 4 – 14 selagi syarat di baris ke 4 dipenuhi
16: endwhile
17: Pilih satu Kunang-Kunang yang miliki keamatanan (i) tertinggi
18: Untuk setiap kelompok yang ada pada Kunang-Kunang yang dipilih
19 Jika ahli kelompok (data) hanya 1, jadikan ia sebagai label/nod dan berhenti.

20 Jika ahli kelompok lebih dari 1
21: Pilih satu data sebagai label/nod (cara caraballo), lalu bagi sisa data menjadi dua kelompok (algoritma pembagi dua sama), kelompokkan masing-masing ahli kelompok kepada dua kelompok baharu.
22: ulangi langkah 19

ALGORITMA 2. Pseudokod AKK-PD

AKK secara semulajadi mempunyai beberapa ciri penting yang sangat berguna untuk melaksanakan tugas pemerolehan (pembelajaran taksonomi). Sifat asas AKK masih dikekalkan untuk menunjukkan bahawa beberapa prinsip asas AKK boleh digunakan dalam pembelajaran taksonomi. Baris ke 8 AKK-PB menunjukkan bagaimana pengelompokan dimulakan. Sentroid yang mempunyai jarak yang kurang dari (p * jarak_terjauh) akan digabung dengan memilih sentroid baharu sebagai kedudukan baharu. Sebagai contoh, andaikan terdapat 50 ekor Kunang-Kunang (populasi). Setiap ekor masing-masing diberi 200 dimensi (*D*). Setiap dimensi, boleh menjadi satu pusat kelompok (sentroid) yang memiliki data sebanyak jumlah kata kerja yang diumpukkan (diekstrak) dari teks. Setiap 200 dimensi (*D*), diukur jaraknya dengan kata nama (*D*) yang wujud dalam kelompok sama ada berdekatan dan perlu dicantumkan ataupun tidak. Nilai p yang digunakan sebagai nilai ambang adalah 5% dari *jarak terjauh*, untuk membetuk $D_{final}$. digunakan 50 populasi, masing-masing terdiri dari 200 (*D*) dimensi asal yang dibangkitkan secara rawak. Masing-masing dimensi, mewakili satu pusat kelompok (sentroid), memiliki data sebanyak jumlah kata nama yang tersedia. Rajah 4.3 menunjukkan contoh pengelompokan dengan 7 data iaitu A, B, D, E, F dan G.




RAJAH 2. Contoh Pengelompokan

Tanda + adalah pusat kelompok (sentroid), dan jika jarak dua atau lebih pusat kelompok kurang dari nilai *p * jarak terjauh* (garis lurus), maka cantumkan dan tentukan nilai tengah (tanda *). Sekiranya nilai $D_{final}$ adalah 5, maka jumlah kelompok yang akan dibentuk adalah 5 kelompok sahaja. Untuk setiap data yang akan dikelompokkan, algoritma akan mencari *d* yang paling kecil di dalam setiap populasi. Nilai ralat terkecil bererti kelompok yang terbentuk adalah kelompok yang lebih baik kualitinya.

Kedudukan baharu ini ditentukan dengan mencari titik tengah di antara semua sentroid yang akan digabungkan. Proses ini berulang untuk setiap kelip sehingga terbentuk beberapa kelompok di dalam setiap kelip. Perlu diingatkan bahawa di dalam AKK-PD, setiap Kunang-Kunang mempunyai satu set solusi kepada masalah. Justeru, setiap ekor Kunang-Kunang asalnya mempunyai satu set sentroid yang jumlahnya ditentukan berdasarkan jumlah data (*kMax*) yang diekstrak dari teks. Baris ke 9 hingga 11 menunjukkan bagaimana proses pengelompokan dilaksanakan. Untuk setiap data (iaitu kata nama) dalam setiap kelip, ukur

jarak (persamaan) dengan semua sentroid di dalam Kunang-Kunang yang sama sahaja. Data dikelompokkan pada sentroid yang terdekat. Kemudian pada baris 12, kualiti setiap kelompok diukur menggunakan fungsi objektif. Fungsi objektif dalam kajian ini adalah dengan mengukur ralat dengan menggunakan persamaan punca kuasa dua hasil tambah ralat. Kelompok yang memiliki ralat yang paling kecil mempunyai keamatan cahaya (i) paling terang. Kini setiap kelip mempunyai nilai i. Populasi (Kunang-Kunang) yang mempunyai nilai fungsi objektif terkecil akan menjadi pusat atau tujuan pergerakan bagi Kunang-Kunang yang lain. Arahan baris ke 4 hingga 14 akan berulang selagi syarat pada baris 4 dipenuhi.
Perbezaan di antara AKK untuk pengelompokan (AKKP) dan AKK untuk pengelompokan berhirarki (AKK-PD) dapat dilihat dari penerangan langkah-langkah pengelompokan berhirarki dalam Jadual 1 berikut:

JADUAL 1. Perbezaan AKKP dan AKK-PD

| | **AKK** | **AKK-PD** |
|---|---|---|
| **Penentuan jumlah kelompok iaitu parameter Kmax** | Jumlah kelompok ditentukan oleh pengguna. | Kmax adalah jumlah kelompok maksimum yang ditentukan secara automatik berdasarkan jumlah kata nama (calon konsep dan tika konsep) yang diperoleh dari teks. Sasaran akhir algoritma AKK-PD adalah setiap kelompok memiliki satu ahli (i.e kata nama) sahaja. |
| **Peranan Sentroid** | AKK menjana satu populasi yang mengandungi sejumlah Kunang-Kunang. Setiap Kunang-Kunang dalam AKK mewakili satu sentroid. | Setiap Kunang-Kunang mempunyai satu set sentroid yang jumlahnya ditentukan berdasarkan jumlah data (kMax) yang diekstrak dari teks. |

Baris ke 17 hingga 20 menunjukkan proses pembinaan hirarki. Baris ke 17 menyatakan bahawa Kunang-Kunang (populasi) yang mengandungi kelompok yang paling berkualiti akan dipilih. Kunang-Kunang yang terpilih adalah Kunang-Kunang yang mempunyai aras keamatan cahaya yang paling terang (i) iaitu jumlah ralat yang paling kecil. Kunang-Kunang tersebut mengandungi sejumlah kelompok. Setiap kelompok mengandungi data (kata nama). Untuk tiap-tiap kelompok, satu daripada data akan dipilih sebagai label atau nod berasaskan algoritma (Caraballo 1999), setelah itu arahan pada baris 18 hingga 21 akan dilaksanakan bagi baki ahli lainnya. Baris 18 ke 21 adalah berasaskan algoritma pembahagi dua sama (Steinbach et al, 2000). Untuk setiap Kunang-Kunang yang mewakili suatu kelompok, akan dibahagikan kepada dua kelompok. Untuk setiap kelompok hasil dari proses pembahagian (di baris 18) akan dibahagi lagi menjadi dua. Proses ini akan diulang kepada setiap kelompok kecuali kelompok yang mempunyai 1 ahli sahaja.

Proses pembahagian ini dipanggil proses *bisecting* (pembahagi dua sama) yang memecah kelompok yang dipilih ke dalam dua kelompok dan menggantikan kelompok asal. Pembahagi dua sama kelompok dijalankan dengan menggunakan asas *K-means* iaitu dengan jumlah *K* (jumlah kelompok baharu) ialah 2. (Cimiano 2006) telah membuktikan bahawa pembahagi dua sama *k-means*, ialah algoritma pembahagi dua sama yang baik dan cepat. Kelebihan menggunakan algoritma pembahagi dua sama adalah ketika proses awal pembahagian atau pemisahan kelompok adalah berasaskan kepada "maklumat global" mengenai objek yang yang hendak dikelompokkan kerana kaedah pengelompokan aglomerat berhierarki melakukan proses pembinaan hierarki tanpa mengambil kira "maklumat global" (Manning et al., 2008). Oleh itu, menurut Manning et al. (2008), algoritma pembahagi dua sama menghasilkan hierarki yang lebih tepat berbanding algoritma aglomerat dalam beberapa aplikasi.

## METODOLOGI KAJIAN

Untuk menguji *kelasakan* atau keteguhan kaedah yang dicadangkan, tiga teks Melayu yang berbeza digunakan untuk memastikan bahawa keputusan yang diperoleh adalah signifikan dan bukannya secara kebetulan. Teks dan set data yang digunakan di dalam kajian ini adalah sama seperti yang digunakan oleh Nazri (2011) yang dibangunkan bersama pakar domain dari setiap domain (iaitu Biokimia, Teknologi Maklumat dan Fekah).

Taksonomi emas setiap domain (teks) dibandingkan dengan kualiti taksonomi yang diperoleh dari set data yang diekstrak menggunakan kaedah pengekstrakan fitur atau ciri yang digunakan oleh Nazri (2011). Nazri (2011) melaporkan bahawa kaedah pengekstrakan ciri yang digunakan beliau adalah berdasarkan kaedah kebergantungan sintaksis Cimiano (2006). Jadual 2 dan 3 memperincikan taksonomi bandingan yang dibangunkan oleh pakar domain berdasarkan kepada teks yang digunakan. Set data yang digunakan oleh Nazri (2011) adalah berasaskan rangka kerja penyelidikan Cimiano (2006). Oleh itu, kualiti taksonomi yang dilaporkan oleh Nazri (2011) adalah hasil dari penggunaan pendekatan Cimiano (2006).

JADUAL 2. Piawaian Emas dan Perbandingan

| | **Teknologi Maklumat** | **Biokimia** | **Fekah** | **Cimiano (2006) Tourism** |
|---|---|---|---|---|
| Bilangan konsep | 478 | 164 | 422 | 293 |
| Bilangan daun | 275 | 113 | 393 | 236 |
| Purata Kedalaman | 2.03 | 3.34 | 2.26 | 3.99 |
| Kedalaman maksimum | 8 | 7 | 6 | 6 |
| Bil. anak maksimum | 96 | 82 | 34 | 21 |
| Purata anak | 4.82 | 7.06 | 5.46 | 5.26 |

Matlamat makalah ini adalah untuk menilai AKK-PD dengan membandingkan keputusan dengan beberapa teknik termaju, supaya generalisasi boleh dibuat berhubung keberkesanan mereka. Sebab utama pendekatan pembelajaran mesin digunakan dalam kajian ini adalah untuk mengatasi isu dan permasalahan yang menyelubungi penggunaan alat pemprosesan bahasa tabie Melayu untuk memeroleh konsep dan membentuk hierarki konsep. Pola bahasa Hearst juga tidak perlu digunakan jika menggunakan pendekatan berasaskan algoritma metaheuristik kerana pola *adalah* tidak semestinya mengandungi hubungan taksonomi. Algoritma yang pernah digunakanuntuk membangunkan taksonomi dari teks Melayu seperti GAHC, CLOSAT dan CLONALG masih bergantung kepada pola bahasa, malahan CLOSAT DAN CLONALG mempunyai terlalu banyak parameter dan tetapan. Berdasarkan sifat algoritma ini yang sukar dilaraskan, kertas ini adalah berkenaan dengan kajian tentang bagaimana untuk membangunkan taksonomi berasaskan AKK secara tanpa selia. Dalam erti kata lain, tugas AKK yang dicadangkan adalah untuk belajar taksonomi dari teks Melayu tanpa bantuan pendekatan linguistik, seperti GAHC dengan Hypernim Oracle (HO).

JADUAL 3. Ringkasan dataset

| | **Teknologi Maklumat** | **Biokimia** | **Fekah** |
|---|---|---|---|
| Jumlah hipernim/hiponim | 119 | 2 | 52 |
| Jumlah ciri kontekstual | 137 | 80 | 162 |
| Bilangan ciri maksimum utk suatu istilah | 61 | 6 | 52 |
| Bilangan ciri minimum utk suatu istilah | 1 | 1 | 1 |
| Min (Jumlah ciri) | 2.08 | 1.51 | 2.47 |
| % istilah dengan 1 ciri sahaja | 69.57% | 76.77% | 62.93% |

Kaedah yang paling praktikal untuk menguji kualiti keluaran AKK-PD adalah dengan membandingkan keputusan AKK-PD dengan teknik yang dibangunkan oleh Cimiano (2006) dan Nazri (2011). Walau bagaimanapun, dalam kajian ini AKK-PD masih dibandingkan dengan kaedah pengelompokan lain termasuk GAHC, HAC dan variansnya untuk analisis perbandingan. Algoritma yang dibandingkan adalah seperti berikut:

1) HAC- Pautan Tunggal.
2) HAC- Pautan Purata.
3) HAC- Pautan Lengkap
4) Algoritma Pembahagi Dua Sama (PDS)
5) Pengelompokan Berhirarki Aglomerat Berpandu (GAHC)
6) Pengelompokan Berpandu Menggunakan AiNet Untuk Pembelajaran Taksonomi (GCAINT).
7) CLOSAT

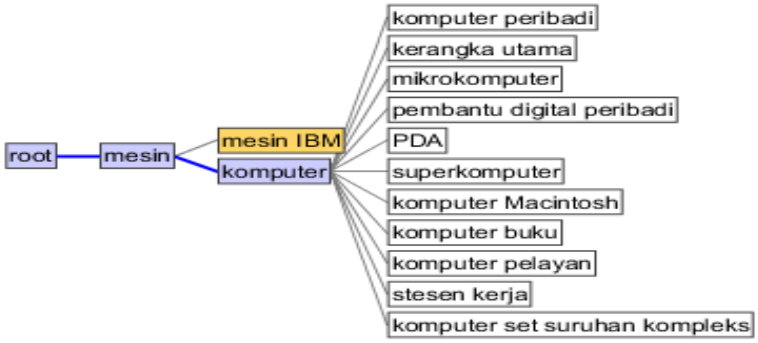
Semua kaedah pengelompokan yang dicadangkan dibangunkan dengan menggunakan bahasa pengaturcaraan Java dan uji kaji dijalankan pada stesen kerja dengan Intel Xeon Processor 3.3 GHz dan 4 GB RAM. Mereka dibandingkan dari segi pertindihan taksonomi, perkaitan leksikal dan mata peratusan.

Ujian kenormalan dijalankan untuk menganalisis keputusan yang diperolehi dalam usaha untuk menentukan sama ada keputusan adalah dalam taburan normal atau tidak. Analisis ini adalah penting untuk membuat keputusan sama ada kaedah statistik boleh digunakan untuk menguji hipotesis dalam eksperimen ini. Analisis varians dijalankan untuk memahami sama ada terdapat perbezaan dalam keputusan di kalangan datase. Dalam erti kata lain, penyelidik ingin menguji sama ada dataset mempunyai pengaruh ke atas prestasi kaedah kelompok yang digunakan dalam kajian ini. Ujian keertian statistik adalah proses yang digunakan untuk menentukan sama ada hipotesis nol boleh ditolak. Hipotesis *null* adalah hipotesis untuk diuji.

Output sistem AKK-PD adalah satu taksonomi yang dipersembahkan dalam XML. XML digunakan untuk mewakili taksonomi dalam usaha untuk membantu penyelidik untuk membandingkan taksonomi rujukan (piawai emas) dan taksonomi yang dijanakan oleh AKK-PD. Perisian SpaceTree yang dibangunkan oleh Plaisant et al. (2002) di University of Maryland digunakan untuk memvisualkan hierarki konsep. Rajah 3 menunjukkan contoh taksonomi diwakili dalam XML manakala Rajah 3 menunjukkan visualisasi taksonomi menggunakan SpaceTree.

```
<?xml version="1.0" encoding="UTF-8" ?>
- <node>
  root
  - <node>
    mesin
    <node>mesin IBM</node>
    - <node>
      komputer
      <node>komputer peribadi</node>
      <node>kerangka utama</node>
      <node>mikrokomputer</node>
      <node>pembantu digital peribadi</node>
      <node>PDA</node>
      <node>superkomputer</node>
      <node>komputer Macintosh</node>
      <node>komputer buku</node>
      <node>komputer pelayan</node>
      <node>stesen kerja</node>
      <node>komputer set suruhan kompleks</node>
    </node>
  </node>
  - <node>
    data
    <node>data mentah</node>
  </node>
```

RAJAH 3. Contoh taksonomi yang diperolehi dari teks Teknologi Maklumat dalam XML




RAJAH 4. Taksonomi Teknologi Maklumat yang dipaparkan oleh SpaceTree

## KEPUTUSAN EKSPERIMEN

Perbandingan pertama yang akan dibincangkan adalah perbandingan prestasi AKK-PD dan Pembahagi Dua Sama kerana AKK-PD adalah hasil hibridisasi AKK dan PDS. Malahan, AKK-PD dan PDS diuji menggunakan set data yang sama. Jadual 4 memaparkan keputusan pengelompokan dan perbandingan keputusan antara AKK-PD dan PDS.

JADUAL 4. Keputusan eksperimen AKK-PD dan PDS

| | **Dataset** | **LR %** | **LP %** | **LF %** | **$P_{TO}$ %** | **$R_{TO}$ %** | **$F_{TO}$ %** |
|---|---|---|---|---|---|---|---|
| AKK-PD | Biokimia | 14.52 | 40.43 | 24.65 | 90 | 15.25 | 30.22 |
| | Fekah | 45.54 | 50.14 | 52.24 | 85 | 16.65 | 27.01 |
| | Teknologi Maklumat | 65.11 | 52.45 | 64.43 | 100 | 17.33 | 32.43 |
| PDS | Biokimia | 12.92 | 35.92 | 14.82 | 72 | 15.25 | 14.73 |
| | Fekah | 32.24 | 39.92 | 34.22 | 65 | 16.65 | 14.86 |
| | Teknologi Maklumat | 47.83 | 44.31 | 35.23 | 60 | 17.33 | 14.04 |

Untuk menentukan prestasi relatif pengelompokan berhierarki AKK-PD, adalah penting untuk membandingkan prestasi AKK-PD dengan algoritma pengelompokan berhierarki tak selia yang lain. Algoritma yang terpilih untuk perbandingan adalah dari 'keluarga' Pengelompokan Berhierarki Aglomerat (HAC) iaitu Pautan-Tunggal, Pautan-Purata dan Pautan-Lengkap. K-Min Pembahagi Dua Sama, Pengelompokan Berhierarki Aglomerat (GAHC), CLOSAT dan GCAINT. Jadual 5 menunjukkan keputusan penilaian antara AK-PD dan lain-lain mengikut set data Fekah, Teknologi Maklumat dan Biokimia.

JADUAL 5. Perbandingan Prestasi ($F_{TO}$) di antara AKK-PD dan Algoritma Bandingan

| | **Fekah (%)** | **Biokimia (%)** | **Teknologi Maklumat (%)** |
|---|---|---|---|
| AKK-PD | 27.01 | 30.22 | 32.43 |
| CLOSAT (Ditala) | 28.41 | 36.25 | 32.28 |
| CLOSAT | 1.59 | 11.76 | 2.17 |
| GCAINT (Ditala) | 28.08 | 25.61 | 31.09 |
| GCAINT | 23.39 | 15.08 | 23.63 |
| GAHC | 23.54 | 10.92 | 24.81 |
| PDS | 14.86 | 14.73 | 14.04 |
| HAC -Tunggal | 14.75 | 12.58 | 14.51 |
| HAC-Purata | 14.47 | 12.52 | 14.51 |
| HAC-Lengkap | 15.09 | 14.15 | 14.63 |

Jadual 6 menunjukkan dengan jelas bahawa prestasi AKK-PD adalah lebih baik berbanding kaedah lain kecuali CLOSAT-PSO dari segi $F_{TO}$ pada semua dataset. CLOSAT-PSO adalah versi CLOSAT yang telah ditingkatkan menggunakan algoritma Pengoptimuman Kerumunan Partikel (PSO) untuk menala parameter CLOSAT. Oleh itu, ia adalah penting untuk menguji sama ada keputusan yang diperolehi tidak berlaku secara kebetulan. Tanggapan asas ujian ini adalah untuk membuktikan bahawa AKK-PD adalah lebih baik daripada kaedah kelompok lain yang sedia ada.

Langkah pertama sebelum memilih kaedah statistik untuk menganalisis keputusan yang diperolehi adalah untuk melaksanakan ujian hipotesis untuk kenormalan. Ujian kenormalan dijalankan untuk menganalisis keputusan yang diperolehi dalam usaha untuk menentukan sama ada keputusan yang diperolehi adalah dalam taburan normal atau tidak untuk menjalan ujian keertian (*significant test*). Persoalannya, pada aras keertian 5%, adakah keputusan memberikan bukti yang mencukupi untuk membuat kesimpulan bahawa $F_{TO}$ diperolehi biasanya diedarkan? Hipotesis nol dan alternatif adalah seperti berikut:

$H_o$: $F_{TO}$ yang diperolehi untuk dataset teragih secara normal.
$H_a$: $F_{TO}$ yang diperolehi untuk dataset tidak teragih secara normal.

Ujian hipotesis dilakukan pada aras keertian 5%, maka alpha = 0.05 . Prosedur statistik Wilk-Shapiro digunakan untuk menguji data untuk kenormalan (Shapiro dan Wilk, 1965). Huruf w digunakan untuk menandakan skor yang normal. Jika nilai w adalah lebih besar daripada 0.05, maka hipotesis nol akan ditolak. Jadual 6 menunjukkan keputusan ujian bagi setiap dataset.

JADUAL 6. Ujian kenormalan

| | **Mean** | **SD** | **Critical Value** | ***w*** | **Result** |
|---|---|---|---|---|---|
| Fekah | 19.30 | 6.12 | 0.803 | 0.7387 | $H_0$ ditolak |
| Teknologi Maklumat | 18.34 | 9.25 | 0.803 | 0.7274 | $H_0$ ditolak |
| Biokima | 20.14 | 6.96 | 0.803 | 0.7813 | $H_0$ ditolak |

Ujian kenormalan berasaskan prosedur statistik Shapiro-Wilk dikira menggunakan aplikasi talian yang dibangunkan oleh (Dittami 2009). Keputusan ujian keertian statistik pada tahap 5%, iaitu, pada aras keertian 5%, data menyediakan bukti yang mencukupi untuk membuat kesimpulan bahawa keputusan berasaskan $F_{TO}$ yang diperolehi tidak berada dalam taburan normal. Oleh itu, tesis ini tidak akan menggunakan ujian berparameter seperti ujian Student t-test atau ANOVA untuk menguji hipotesis. Oleh yang demikian, ujian tidak-berparameter seperti Wilcoxon Rank Sum Test akan digunakan untuk menguji hipotesis kajian ini.

## UJIAN KEERTIAN

Pada aras keertian 5%, adakah keputusan (data) yang diperolehi memberikan bukti yang mencukupi untuk membuat kesimpulan bahawa keputusan $F_{TO}$ bagi setiap kaedah kelompok bandingan kajian ini melebihi keputusan $F_{TO}$ yang dihasilkan oleh AKK-PD. Dalam usaha untuk menjawab soalan ini, ujian bukan parametrik Wilcoxon Rank Sum digunakan bagi menguji keertian keputusan ini. Hipotesis nol dan alternatif adalah seperti berikut:

$H_0$: $F_{TO}$ kaedah pengelompokan *x* adalah sama atau sama dengan $F_{TO}$ AKK-PD.
$H_a$: $F_{TO}$ kaedah pengelompokan *x* adalah lebih kecil atau sama dengan $F_{TO}$ AKK-PD.

Ujian hipotesis ini ialah hujung kiri kerana tanda lebih kecil digunakan (<) dalam hipotesis alternatif. Ujian hipotesis ini dijalankan pada aras keertian 5% atau 0.05. Merujuk kepada Jadual Nilai Kritikal ujian Wilcoxon Rank Sum, nilai kritikal T = 6. Mengikut prosedur, didapati bahawa T = 6. $H_0$ ditolak jika nilai ujian statistik adalah lebih kecil atau sama dari nilai T yang diambil dari jadual nilai kritikal ujian Wilcoxon Rank Sum (Swed & Eisenhart, 1943). Jadual 7 memaparkan keputusan ujian statistik.

JADUAL 7. Perbandingan Prestasi ($F_{TO}$) di antara AKK-PD dan Algoritma Bandingan

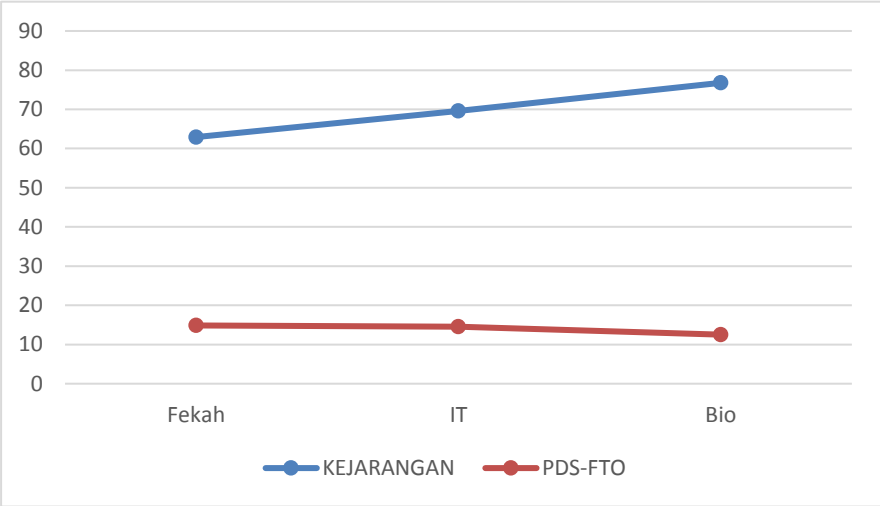
| | **Fekah (%)** | **Biokimia (%)** | **Teknologi Maklumat (%)** | ***M*** | **Ujian Statistik** | **Tolak $H_0$** |
|---|---|---|---|---|---|---|
| AKK-PD | 27.01 | 30.22 | 32.43 | | | |
| CLOSAT (Ditala) | 28.41 | 36.25 | 32.28 | 6 | 12 | Tidak |
| CLOSAT | 1.59 | 11.76 | 2.17 | 6 | 6 | Ya |
| GCAINT (Ditala) | 28.08 | 25.61 | 31.09 | 6 | 6 | Ya |
| GCAINT | 23.39 | 15.08 | 23.63 | 6 | 6 | Ya |
| GAHC | 23.54 | 10.92 | 24.81 | 6 | 6 | Ya |
| PDS | 14.86 | 14.73 | 14.04 | 6 | 6 | Ya |
| HAC -Tunggal | 14.75 | 12.58 | 14.51 | 6 | 6 | Ya |
| HAC-Purata | 14.47 | 12.52 | 14.51 | 6 | 6 | Ya |
| HAC-Lengkap | 15.09 | 14.15 | 14.63 | 6 | 6 | Ya |

## PERBINCANGAN

Ketiga-tiga set data yang digunakan mengalami masalah kejarangan data yang tinggi kerana lebih 60% daripada istilah (kata nama) di dalam set data tersebut hanya memiliki satu ciri sahaja.. Ciri yang diperolehi dari teks menggunakan kaedah keberantungan sintaksis jelas menunjukkan masalah kejarangan data yang serius. Kajian ini berusaha membangunkan algoritma yang lasak apabila berdepan dengan isu ini. Jadual 6 memaparkan keputusan yang diperolehi oleh AKK-PD dan PDS. Ujian statistik dilakukan pada aras keertian 5% dan ujian adalah hujung kanan, di mana alpha = 0.05 dijalankan bagi membuktikan bahawa algoritma hybrid adalah lebih berkesan berbanding PDS yang tidak dihybrid dengan AKK. Menggunakan ujian bukan parametric, kawasan di sebelah kiri titik kritikal $T_L = 5$ manakala $T_U = 16$ justeru $H_0$ akan ditolak jika nilai cerapan T adalah lebih kecil atau sama dengan nilai $T_L = 5$ atau lebih besar atau sama dengan $T_U = 16$. Nilai cerapan hipotesis nol bagi keputusan menunjukkan T = 6. Oleh itu, nilai kritikal T jatuh di rantau penolakan, lantas $H_0$ adalah ditolak maka kesimpulan dibuat adalah bahawa keputusan ($F_{TO}$) *meningkat dengan signifikan* apabila menggunakan AKK-PD berbanding PDS. Oleh kerana aras keertian yang digunakan dalam ujian ini adalah bersamaan 0.05, maka dapatlah disimpulkan bahawa AKK-PD telah meningkatkan $F_{TO}$ dengan ketara pada semua set data.

Jadual 7 menunjukkan keputusan ujian statistik yang signifikan iaitu pada tahap 5%. Jadual 7 memaparkan keputusan atau prestasi tujuh (7) algoritma pengelompokan berhirarki yang diperolehi oleh Nazri (2011) digunakan sebagai asas perbandingan dengan keputusan yang diperolehi oleh AK-PD bila diuji pada teks yang sama. Data yang dikumpulkan menyediakan bukti yang cukup bahawa populasi $F_{TO}$ bagi setiap kaedah kelompok didapati lebih rendah daripada min $F_{TO}$ AKK-PD kecuali CLOSAT-PSO. Namun, AKK-PD secara purata menghasilkan taksonomi yang lebih baik daripada kaedah kelompok lain yang digunakan dalam kajian ini. Jadual 8 di atas menunjukkan bahawa AKK-PD mendapat keputusan lebih baik pada data Teknologi Maklumat berbanding CLOSAT tetapi lebih rendah di set data Fekah dan Biokimia berbanding CLOSAT. Perbandingan ini menunjukkan AKK-PD mempunyai prestasi yang memberansangkan dan mempunyai potensi tinggi. Keputusan yang dipaparkan adalah hasil AKK-PD yang tidak doptimumkan sedangkan

CLOSAT telah ditala dan dioptimumkan menggunakan PSO. Adalah penting untuk menentukan sama ada algoritma cadangan iaitu AKK-PD telah meningkatkan $F_{TO}$ untuk semua dataset dengan ketara sekali atau tidak. Satu ujian statistik dijalankan untuk menguji sama ada taburan keputusan ($F_{TO}$) yang diperolehi dari menggunakan AKK-PD berada disebalah kanan dari taburan $F_{TO}$ yang diperolehi dari menggunakan PDS. Oleh itu, hipotesis nol adalah, kedua-dua agihan populasi $F_{TO}$ bagi kedua-dua algoritma adalah sama. Ujian Wilcoxon Rank Sum digunakan kerana taburan data yang ada tidak normal. Dalam usaha untuk menguji hipotesis yang dibincangkan dalam bahagian ini, diandaikan bahawa dua populasi mempunyai bentuk taburan yang sama. Menggunakan aras keertian 5%, hipotesis nol telah dpat ditolak dan kajian menyimpulkan bahawa $F_{TO}$ AKK-PD telah berjaya ditingkatkan dengan ketara. Rajah 5 menunjukkan graf peratus kejarangan melawan keputusan $F_{TO}$ bagi PDS. Hal ini menunjukkan semakin tinggi peratus kejarangan, semakin rendah $F_{TO}$ bagi algoritma PDS.




RAJAH 5. Graf perbandingan peratus kejarangan dan $F_{TO}$ PDS

Analisis varians dijalankan dalam usaha untuk menjawab soalan berikut:
*Pada aras keertian 5%, adakah set data (keputusan) terkumpul memberikan bukti yang mencukupi untuk membuat kesimpulan bahawa wujud perbezaan dalam keputusan yang diperolehi dari ketiga-tiga dataset (Fekah, Biokimia dan Teknologi Maklumat) yang diuji pada semua algoritma?*

Analisis varians dijalankan berdasarkan data di dalam Jadual 8 yang memaparkan perbandingan $F_{TO}$ antara teknik AKK-PD dengan lain-lain algoritma pada kesemua set data.

JADUAL 8. Perbandingan prestasi AKK-PD dengan kaedah lain pada setiap set data

| | **Biokimia** | **Fekah** | **Teknologi Maklumat** |
|---|---|---|---|
| HAC-Tunggal | 12.58 | 14.75 | 14.51 |
| HAC-Purata | 12.52 | 14.47 | 14.51 |
| HAC-Lengkap | 14.15 | 15.09 | 14.63 |
| K-Min-PD | 14.73 | 14.86 | 14.04 |
| GAHC | 10.92 | 23.54 | 24.81 |
| GCAINT-PSO | 25.61 | 28.08 | 31.09 |
| CLOSAT-PSO | 36.25 | 28.41 | 32.28 |
| AKK-PD | 30.22 | 27.01 | 32.43 |

Ujian statistik ini dijalankan dengan matlamat untuk membuktikan bahawa kaedah cadangan adalah lebih kukuh atau lasak terhadap isu kejarangan data. Oleh kerana data (keputusan) yang diperolehi bukan dalam taburan normal, ujian Kruskal-Wallis digunakan

untuk menguji tanggapan asas bahawa sifat dataset tidak menjejaskan $F_{TO}$ Oleh yang demikian, hipotesis nol adalah seperti berikut:

> $H_0$: Taburan keputusan ($F_{TO}$) bagi setiap set data adalah dengan median yang sama.
> $H_a$: Taburan keputusan ($F_{TO}$) bagi setiap set data adalah dengan median yang tidak sama.

Aras ujian keertian Kruskal-Wallis adalah $\alpha$=0.05, dan darjah kebebasan adalah df=5–1=4. Oleh kerana saiz sampel adalah kurang dari 5 maka anggaran normal tidak boleh digunakan maka taburan Chi-Square digunakan dan nilai kritikal H adalah Hc=8.333. Oleh itu, rantau penolakan untuk ujian ini adalah R={H≥8.333}.
Statistik H dikira berdasarkan rumus berikut :

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k}\right) - 3(N+1)$$

$$= \frac{12}{15(15+1)}\left(\frac{9^2}{3} + \frac{12^2}{3} + \frac{30^2}{3} + \frac{34.5^2}{3} + \frac{34.5^2}{3}\right) - 3(15+1) = 10.425$$

Oleh kerana H=10.425≥8.333, maka dapatlah diputuskan bahawa hipotesis null ditolak justeru dapatlah disimpulkan bahawa pada aras keertian 0.05, terdapat bukti yang mencukupi untuk menyatakan bahawa tidak semua median populasi (keputusan) adalah sama. Maka, ini menunjukkan bahawa data yang berbeza lanskapnya mempunyai kesan kepada prestasi algoritma. Dalam erti kata lain, dapatlah disimpulkan bahawa set data mempengaruhi prestasi kaedah (algoritma) pengelompokan dan keputusan yang diperolehi adalah disebabkan oleh sifat data.

## KESIMPULAN

Pembangunan taksonomi untuk Web dan Internet secara manual memerlukan masa yang panjang untuk dibangunkan dan sekali gus meningkatkan kos. Justeru kaedah kecerdasan buatan digunakan oleh saintis computer untuk membangunkan taksonomi secara automatik dari teks. Teks berbahasa Melayu sangat jarang digunakan sebagai asas pembentukan taksonomi Web justeru kajian ini bertujuan untuk meneroka keberkesanan algoritma metaheuristik dalam membina hierarki konsep atau taksonomi dari Teks Melayu. Algoritma metaheuristik yang dicadangkan adalah berasaskan konsep keamatan cahaya serangga Kunang-Kunang. Konsep atau sifat serangga ini yang telah dijadikan algoritma pengkomputeran yang dikenali sebagai algoritma Kunang-Kunang (AKK) telah dihibrid dengan algoritma Pembahagi Dua Sama (PDS). AKK yang diubahsuai ini dipanggil Algoritma Kunang-Kunang Pembahagi Dua Sama (AKK-PD). Kajian kesusasteraan menunjukkan AKK julung-julung kali diubahsuai untuk pembelajaran taksonomi. Tiga teks Melayu telah digunakan dalam kajian ini iaitu teks dari bidang Fekah, Teknologi MAklumat dan Biokimia. Keputusan yang diperolehi menunjukkan prestasi AKK-PD telah mengatasi kesemua algoritma bandingan lain dalam perbandingan kualiti. Walau bagaimanapun, AKK-PD tidak mampu menandingi CLOSAT-PSO kerana CLOSAT-PSO telah ditala parameternya dengan menggunakan algoritma Pengoptimuman Kerumunan Partikel (PSO). PSO telah digunakan untuk mencari parameter yang optimum dalam meningkatkan prestasi CLOSAT manakala AKK-PD tidak ditala dengan menggunakan algoritma metaheusristik PSO. Perbandingan sebelum ini diantara AKK-PD dengan CLOSAT (tanpa PSO), AKK-PD mempunyai prestasi yang lebih baik. Walau bagaimanapun, prestasi CLOSAT-PSO masih

boleh ditandingi dengan membuat beberapa penambahbaikan kepada AKK kerana PDS sendiri mempunyai kekurangan yang tersendiri. AKK-PD dilihat sebagai kaedah yang berpotensi untuk digunakan dalam pembelajaran taksonomi dan juga sebagai algoritma pengelompokan berhierarki yang berkesan. Keputusan menunjukkan bahawa prestasi AKK-PD adalah lebih baik berbanding dengan algoritma PDS pada ketiga-tiga set data. Ukuran yang digunakan di dalam perbandingan dalam jadual di atas tersebut adalah Ukuran F-Leksikal (LF) dan F-Tindanan Taksonomi ($F_{TO}$). Prestasi AK-PD adalah konsisten pada ketiga-tiga set data dan ini menunjukkan bahawa AKK-PD sangat sesuai untuk digunakan dalam pembelajaran taksonomi walaupun set data mengalami masalah kejarangan data yang sangat serius. Keputusan yang menunjukkan bahawa prinsip biologi boleh membawa kepada pembangunan alat pembelajaran taksonomi yang lebih berkesan dari beberapa algoritma bandingan seperti CLOSAT dan CLONALG.

**PENGHARGAAN**

Penulis ingin mengucapkan terima kasih kepada sidang editor GEMA yang telah memberikan nasihat dan panduan berharga untuk menghasilkan makalah ini. Kajian ini telah dijalankan dengan bantuan Geran Galakan Penyelidik Muda Universiti Kebangsaan Malaysia (GGPM-2012-001).

## RUJUKAN

Abhay Jain, Srujan Chinta & Tripathy, B.K. (2017). Stabilizing Rough Sets Based Clustering Algorithms Using Firefly Algorithm over Image Datasets. International Conference on Information and Communication Technology for Intelligent Systems Conference Proceedings, 325-332. Springer, Cham.

Abhishek Bafna & Wiens, J. (2015). Automated feature learning: Mining unstructured data for useful abstractions. 2015 IEEE International Conference on Data Mining Conference Proceeding, 703-708.

Amirah Ismail, Joy, M.S., Sinclair, J.E. & Mohd Isa Hamzah. (2009). A metametadata taxonomy to support semantic searching algorithms in metadata repository. International Conference on Electrical Engineering and Informatics Conference Proceedings, vol. 2, 1-6. IEEE.

Charles W.G. (2000). Contextual correlates of meaning. Applied Psycholinguistics 21, 505-524.

Cimiano, P., Hotho, A. & Staab, S., (2005). Learning Concept Hierarchies From Text Corpora Using Formal Concept Analysis. J. Artif. Intell. *Res.(JAIR). Vol. 24*(1), 305-339.

Cimiano, P. & Staab, S. (2005).Learning Concept Hierarchies from Text with a Guided Agglomerative Clustering Algorithm. International Conference on Machine Learning 2005 (ICML 2005) Conference Proceedings, Bonn Germany.

Cimiano, P. (2006). *Ontology Learning and Population From Text*. Springer Berlin.

De Castro, L.N. & Timmis, J. (2002). *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer Science & Business Media.

de Mantaras, R.L. & Saitia, L. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. 16th European Conference on Artificial Intelligence Conference Proceedings, Vol. 110, 435. IOS Press.

Fister, I., Fister Jr, I., Yang, X.S. & Brest, J. (2013). A Comprehensive Review of Firefly Algorithms. *Swarm and Evolutionary Computation. Vol. 13*, 34-46.

Firth, J.R. (1957). *A Synopsis of Linguistic Theory 1930-1955*. Longman. London.

Izfa Riza Hazmi, & Sharifah Aliya Syed Sagaff (2018). Fireflies Population and the Aquaculture Industry (Coleoptera: Lampyridae) of the Sungai Sepetang, Kampung Dew, Perak, Malaysia. *Serangga. Vol. 22*(2).

Harris, Z. (1954). Distributional Structure. *Word. Vol. 10*(23), 146-162.

Harris, Z. (1968). *Mathematical Structure of Language*. Wiley.

Herna Banati & Monika Bajaj. (2013). Performance Analysis of Firefly Algorithm for Data Clustering Int. J. *Swarm Intelligence. Vol. 1*(1).

Jay J. Jiang & Conrath D.W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. International Conference Research on Computational Linguistics ROCLING X Conference Provceedings Taipei, Taiwan, 1997.

Lefever, E. (2016). A Hybrid Approach to Domain-independent Taxonomy Learning. *Applied Ontology. Vol. 11*(3), 255-278.

Lewis, S.M. & Cratsley, C.K. (2008). Flash Signal Evolution, Mate Choice, and Predation in Fireflies. *Annual Review of Entomology. Vol. 53*, 293-321

Luu Anh Tuan, Yi Tay, Siu Cheung Hui & See Kiong Ng. (2016). Learning term embeddings for taxonomic relation identification using dynamic weighting neural network. Conference on Empirical Methods in Natural Language Processing Conference Proceedings, 403-413.

Miller, G.A. & Charles, W.G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes. Vol. 6*(1),1-28.

Mohammed, A.J., Yusof, Y. & Husni, H. (2014). Weight-based Firefly algorithm for document clustering. First International Conference on Advanced Data and Information Engineering (DaEng-2013) Conference Proceedings, 259-266. Singapore.

Mohd Zakree Ahmad Nazri, Siti Mariyam Shamsuddin, Azuraliza Abu Bakar & Tarmizi Abd Ghani. (2008). Using linguistic patterns in FCA-based approach for automatic acquisition of taxonomies from Malay text. 2008 International Symposium on Information Technology Conference Proceedings, Vol. 2, 1-7.

Mohd Zakree Ahmad Nazri, Siti Mariyam Shamsuddin, Azuraliza Abu Bakar & Salwani Abdullah. (2011). A Hybrid Approach for Learning Concept Hierarchy From Malay Text Using Artificial Immune Network. *Natural Computing. Vol. 10*, 275-304.

Nur Hudawiyah, O., Nurul Wahida & S. Norela. (2015, September). Gross anatomy of central nervous system in firefly, Pteroptyx tener (Coleoptera: Lampyridae). In *AIP Conference Proceedings* (Vol. 1678, No. 1, p. 020017). AIP Publishing.

Nayak, J., Nanda, M., Nayak, K., Naik, B. & Behera, H.S. (2014). *An Improved Firefly Fuzzy C-means (Fafcm) Algorithm for Clustering Real World Data Sets. In Advanced Computing, Networking and Informatics*. Springer, Cham.

Ristoski, P., Faralli, S., Ponzetto, S.P. & Paulheim, H. (2017). August. Large-scale taxonomy induction using entity and word embeddings. The International Conference on Web Intelligence Conference Proceedings, 81-87.

Sarma, P.N. & Gopi, M. (2014). *Energy Efficient Clustering Using Jumper Firefly Algorithm in Wireless Sensor Networks*. arXiv preprint arXiv:1405.1818.

Senthilnath, J., Omkar, S.N. & Mani, V. (2011). Clustering Using Firefly Algorithm: Performance Study. *Swarm and Evolutionary Computation. Vol. 1*(3), 164-171.

Wan Faridah Akmal Jusoh, Nor Faridah Hashim & Nur Azura Adam. (2013). *Distribution of the Synchronous Flashing Beetle, Pteroptyx Tener Olivier (Coleoptera: Lampyridae), in Malaysia*. The Coleopterists Bulletin.

Wan Juliana, W.A, Md. Shahril, M.H., Nik Abdul Rahman, N.A., Nurhanim, M.N., Maimon Abdullah, M. & Norela Sulaiman. (2012). Vegetation Profile of the Firefly Habitat Along the Riparian Zones of Sungai Selangor at Kampung Kuantan, Kuala Selangor. *Malaysian Applied Biology. Vol. 41*(1), 55-58.

Wang, C., He, X. & Zhou, A. (2017). A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances. *2017 Conference on Empirical Methods in Natural Language Processing* Conference Proceedings, 1190-1203.

Wang, C., Fan, Y., He, X. & Zhou, A. (2018). Predicting Hypernym–hyponym Relations for Chinese Taxonomy Learning. *Knowledge and Information Systems*. 1-26.

Wong, L.A., Shareef, H., Mohamed, A. & Ibrahim, A.A. (2014). Optimal battery sizing in photovoltaic based distributed generation using enhanced opposition-based firefly algorithm for voltage rise mitigation. The Scientific World Journal.

Xiujuan Lei, Fei Wang, Fang-Xiang Wu, Aidong Zhang & Pedrycz, W. (2016). Protein Complex Identification Through Markov Clustering With Firefly Algorithm on Dynamic Protein–protein Interaction Networks. *Information Sciences. Vol. 329*, 303-316.

Yang, X.S. (2008). *Nature-Inspired Metaheuristic Algorithm*. Frome.

Yang, X.S. (2010). Firefly Algorithm, Stochastic Test Functions and Design Optimisation. *International Journal of Bio-Inspired Computation. Vol. 2*(2), 78-84.

Yong-Bin Kang, Haghigh, P.D. & Burstein, F., (2016). Taxof Inder: a Graph-based Approach for Taxonomy Learning. *IEEE Transactions on Knowledge and Data Engineering. Vol. 28*(2), 524-536.

Zipf, G.K. (1935.) *The Psychobiology of Language*. Houghton-Mifflin.

## PENULIS

Mohd Zakree Ahmad Nazri (Ph.D) ialah Profesor Madya di Pusat Teknologi Kecerdasan Buatan, Fakulti Teknologi dan Sains Maklumat (FTSM), UKM. Bidang pengkhususan dan kepakaran beliau ialah Sistem Sokongan Keputusan dan Analitik Bisnes. Kajian yang sedang dijalankan adalah pembangunan algoritma pembelajaran mesin dan permodelan keputusan untuk sokongan keputusan.

Tri Basuki Kurniawan adalah pelajar Program Doktor Falsafah dalam bidang perlombongan data di FTSM UKM. Beliau berasal dari Palembang, Indonesia. Beliau juga ada bekas pensyarah universiti di Palembang.

Abdul Razak Hamdan (Ph.D) adalah seorang Profesor dalam bidang kecerdasan buatan di FTSM UKM. Bidang penyelidikan adalah merentas bidang di antaranya dalam bidang sistem maklumat, polisi dan kajian strategik dan pembelajaran mesin.

Salwani Abdullah (Ph.D) adalah Profesor dalam bidang analitik preskriptif (pengoptimuman). Bertugas di FTSM UKM dengan pengkhususan dan kepakaran dalam bidang pengautomasian perancangan dan penjadualan terutamanya dalam pembangunan algoritma diinspirasi alam.

Mohammed Azlan Mis (Ph.D) ialah seorang pensyarah di Pusat Pengajian Bahasa dan Linguistik, UKM. Bidang pengkhususan dan kepakaran beliau ialah Sosiolinguistik dan Dialektologi. Dalam bidang penyelidikan, kajian-kajian yang sedang dilakukan berkenaan dengan pemilihan bahasa di sempadan, doktor-pesakit dan juga dalam sektor pelancongan.